Personality Prediction

Vishnu Vardhan Reddy Yeruva
(Software Engineering)
San Jose State University
San Jose, Santa Clara
vishnuvardhanreddy.yeruva@sjsu.edu

Khushil Ketankumar Modi
(Software Engineering)
San Jose State University
San Jose, Santa Clara
khushilketankumar.modi@sjsu.edu

Nevil Viplav Bahi Shah
(Software Engineering)
San Jose State University
San Jose, Santa Clara
nevilviplavbhai.shah@sjsu.edu

Abstract—We have researched, designed and developed an application for personality prediction which will play a pivotal role in developing one's behavior. Personality prediction ensures different types of behaviors are categorized into 5 base types and predicts a score for each category that compares with the mean of each category and suggests the perfect areas to be worked to make him balanced. We have picked a finest regression and ML algorithms that gives the best accuracy of 99% after downsampling and upsampling the data after iterating multiple times in the muller loop

I. Introduction:

Nowadays, more emphasis and research is given on the ability of a person's character. For example, in the IT field, behavioral assessment will be a required filter incase of emergencies and stressful situations, how stable or balanced he can keep himself. Hence, it is very important to keep track of a person's conditions and on which aspects he has to work and focus on is definitely needed.

This prediction model's main aim is to concentrate on the features that a person is lacking on an average on his complete behavior and to make him balanced.

II. Approach

Everytime to find a solution we have to divide a problem in its simplest possible ways. We completely approached this problem in agile completely for any simple step that we have divided from categories and defining golden clusters and extracting features out of it. There is a lot of brainstorming and implementing many possible solutions differently and picking the best possible one out of it to integrate it with our project.

III. Questions:

- What is the importance of behavioral assessments?
- Why are people leaving industries and what is their after life?
- What are the challenges that we will be facing technically?
- What sort of output makes it more clean, understandable and effective?
- What algorithms are required and what changes to be made accordingly?

IV. Experiments:

- Researched on several behaviors and what made them to struggle
- Made some bases on algorithms and experimented with various features
- Added some miscellaneous features, after trails getting failed dropped them off
- Contributed heavily to research on dataset

V. Team name and members:

Team Dragonites	
Team Members	**Roles**
Khushil Modi (015923115)	Cleaning, Dimensionality Reduction, GMM, Fractal, Iteration - 1, SMOTE, upsampling
Nevil Shah (015964975)	Categorizing, KMeans, DBScan, Fractal, Iteration - 2, labeling/rating,

	classification, downsampling
Vishnu Yeruva (015917551)	Data & Amalgamation, Silhouette, Birch, Fractal, Iteration - 3, Latent Features, Interactive Slider, muller loop

VI. Data

Gone for most preferable and reliable sources - KAGGLE

DataSet-I

It is a five-factor model (FFM) [1] which is a taxonomy, or grouping, for personality traits. When factor analysis (a statistical technique) is applied to personality survey data, some words used to describe aspects of personality are often applied to the same person.

Data set 2

This helps in 5-point Likert scale. (1: Strongly Disagree, 5: Strongly Agree) [2]. Based on the assigned metric and on this assigned condition will be derived a condition which is oddly imbalanced.

Data set 3

This is amalgamation of Data Set 1, Data Set 2 and Dataset 3 (which has been scraped)

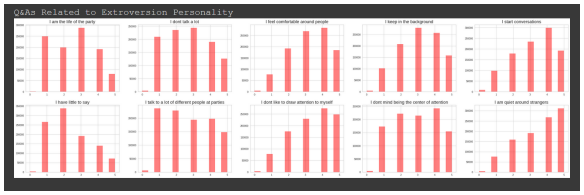
Discussion on Amalgamation of the Datasets

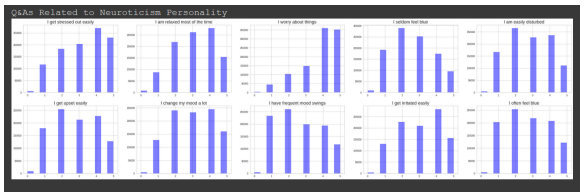
Data Set 1 appropriately provides what we want exactly but to group things near as a cluster we need something which is more organized and better in order to improve efficiency for sse shoulder from 91% to 99% we had to move to data2 and data3. Later to extract latent features and manifolds we appended all the researched required columns from 3 data sets to single and added those features to.

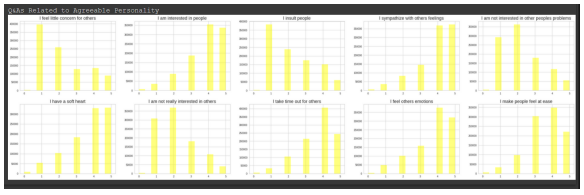
VII. Exploratory Data Analysis and Visualization

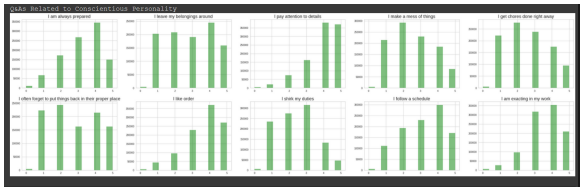
Categorization

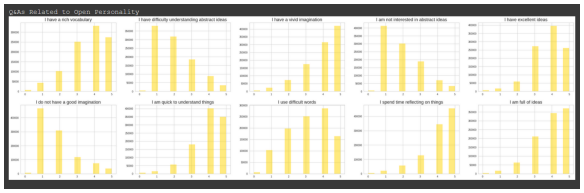
Categorization of different personalities into 5 base aspects and training with sample questions for each category that again picks up a baseline on how the questions is going to be patterned.









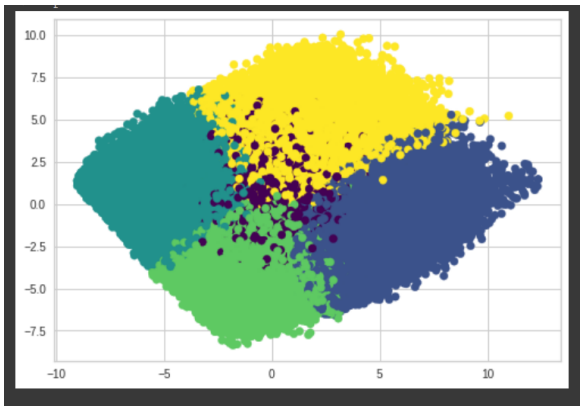


VIII. Clustering

Different clustering techniques

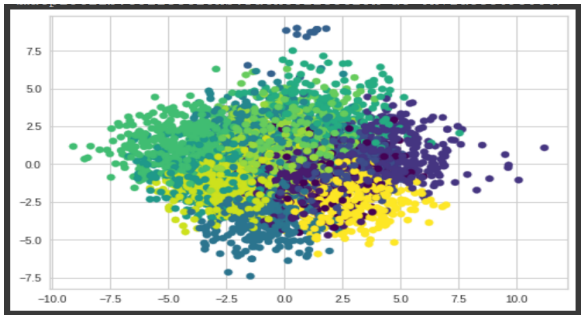
K-Means

Visualizing 5 categories as 5 clusters using K-Means with PyTorch [3].



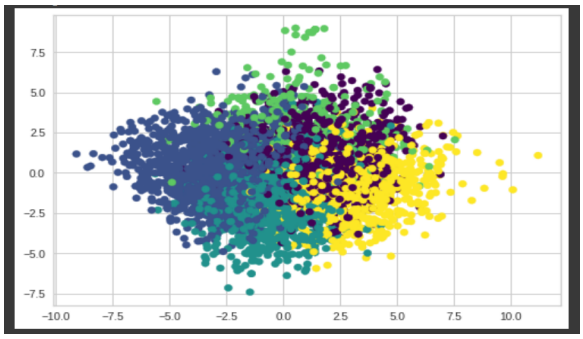
GMM

Visualizing 5 categories as 5 clusters using GMM with DBScan.



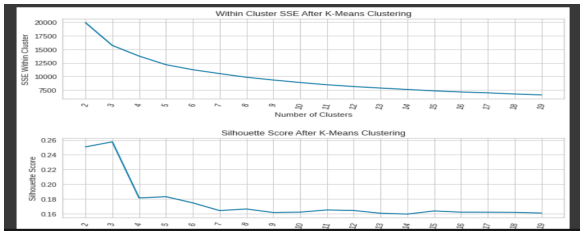
Birch

Visualizing 5 categories as 5 clusters using BIRCH[7].



Fractal Clustering

After looking at the performance of various clusters using K-Means, performance is evaluated within cluster SSE and silhouette score by using robust scaling so that centering and scaling are therefore not influenced by a small number of very large marginal outliers as they are based on percentiles. After that applied K-Means clustering with the optimal number of clusters defined [6].



Golden Cluster

What is the golden cluster?

A. Getting the better cluster out of a different cluster that is having all the features that every other cluster had.

We have iterated on different trails to achieve as low as possible and preferable silhouette score.

First Trail

```
clustering performance
----------------------------------
silhouette score: 0.26
sse withing cluster: 15730
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:16
A value is trying to be set on a copy of a slice from a DataFra
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org
  app.launch_new_instance()
```

Second Trail

```
clustering performance
----------------------------------
silhouette score: 0.19
sse withing cluster: 8426
```

Final Trail

```
clustering performance
----------------------------------
silhouette score: 0.17
sse withing cluster: 1547
```

Performed iterations until desired score is achieved [5].

What is the objective function?

A. Taking the resultant model and evaluating it against the training dataset.

IX. Latent Variables and Manifolds

What are the latent variables you added in and how did that enhance your metrics?

We summed up the whole values from each person into a ‘personality’ column that helped us in labeling appropriate sets for values.

We made our project much more understandable and easier by labeling the values in personality into 3 different categories which is used to turn into a ‘Test’ column where we achieved our results on comparing the level of maturity that a person is having for a particular behavior.

We made a mean for each person's scores on their behavior i.e.’Avg_Personality’ column that helped in achieving down_sampling, over_sampling and results finally.

X. Classification

What metrics are you using ? [f1, precision recall, etc]
We computed all the values into 3 discrete sets labeling the behavior on a likert scale of ‘Good’, ‘Bad’, ‘Medium’

0	Medium
1	Good
2	Bad
3	Medium
4	Medium
...	...
1829	Medium
1830	Medium
1831	Good
1832	Bad
1833	Good

Muller loop

We Used KNN for iteration with 90% of accuracy in 0.02 seconds.

XI. Regression

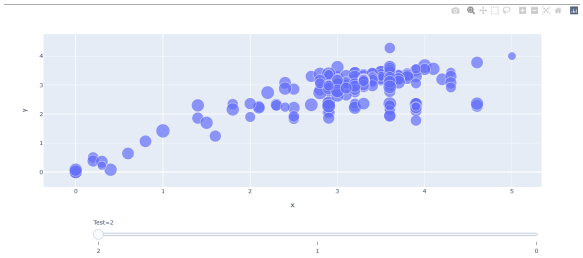
[Answer to question 1]

We are able align all those values in a column in a more organized way and very efficiently using RandForest

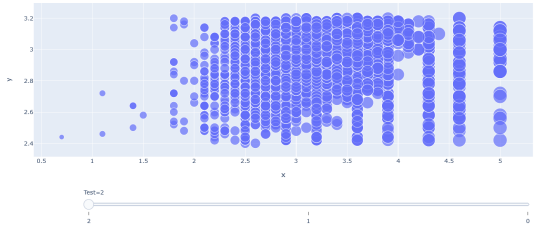
Muller Loop:
We Used Random Forest regressor for iteration with 100% of accuracy in 0.04 seconds.

XII. Distributions of your Data

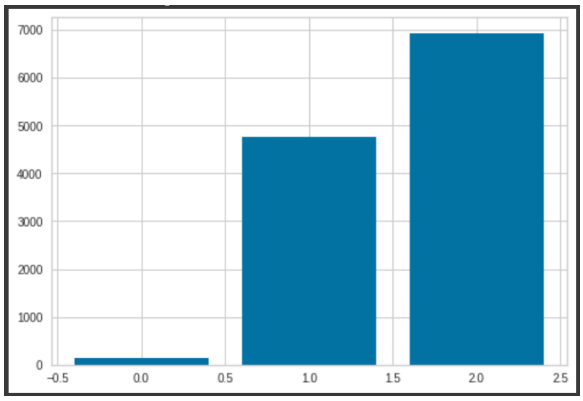
EDAV
After downsampling and oversampling the representation of features with the ratings at its best compared to average behavior of that person is visualized. From this we can grab the personality on what he can seriously work by comparing every dashboard that has been visualized in the colab.



For Down-Sample Data (Openness)



For Over-Sample Data (Openness)
While ranking the categories we were able to find out the frequency distribution over 3 parameters we set [4].



Selected Features
Which features did you select?

- We have added some miscellaneous features as to what type of movies are being watched by users which is allowed by the 2nd dataset that we can judge one's behavior on how frequent he is watching a genre.
- With classification and regression labeled the personality behavior in likert scale.

Algorithms used to select features

And why? [selected because these algos told me these were the top features!!]

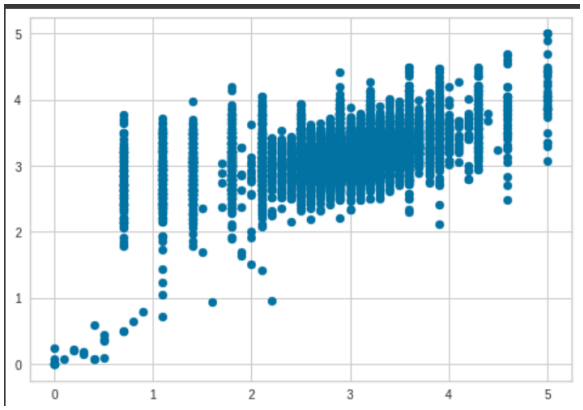
From the start we have been very clear on our idea on how to proceed. But moving forward we have to make some changes on the way to reach the result.

- So, on the first feature we took an advantage of dataset
- On the 2nd feature we are clear on the feature to be extracted, so on the way of selecting algorithms we changed the output of our feature into string based into 3 discrete sets

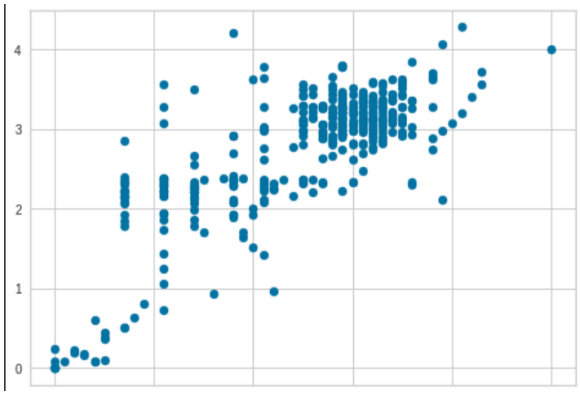
Changing Data Distributions

What distribution of data will give you a better model?

A. Barplots



Visualizing before sampling
(Extraversion)



Visualizing after down-sampling
(Extraversion)

Before sampling the data is very uneven to represent accurately to show the appropriate behavior. We achieved the balanced data to find which category is lacking in rating.

What are the worst and best distributions of your datasets? [Why?]

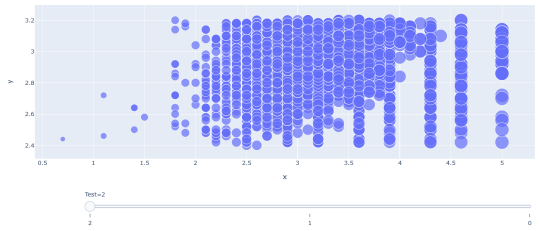
Before sampling is the worst distribution, because every value is uneven that leads nowhere. Sampling made everything organized and appropriate to compare.

XII. Data Narrative and Conclusions

We have grown on how we think while making this. We started of after a strong and solid technology stack and a proper idea by knowing what to achieve from this architecture, but later being

faced with many challenges our approach towards the solution had changed then we researched on the algorithms that on what is needed for features and how the algorithm affects our program on showing the results, so what changes should be made to achieve the results.

Personality Prediction is a dream project not only for us, but it is a kind of relief for people who have been waiting for a bounce back in life. There are people who had been into depression and could not find what went wrong with their behavior previously and now. So, we thought of creating something to help them find their innerselves and what can be improved further to show the stability in their actions as well as words.



If we are able to concentrate on the image, we are clearly comparing the person's complete personality's mean with the personality that he's lacking in behind with. Where as 0,1,2 represents good, medium and bad respectively

	Unnamed: 0	openness	agreeableness	neuroticism	conscientiousness	extraversion	Personality	Test	Avg_Personality
0	0	3.6	1.4	2.1	1.8	4.6	13.5	2	2.70
1	1	5.0	2.9	4.3	3.9	2.9	19.0	1	3.80
2	2	2.9	2.1	3.2	1.4	1.8	11.4	0	2.28
3	3	3.9	3.9	2.9	3.2	2.9	16.8	1	3.36
4	4	3.9	3.9	2.5	3.2	1.8	15.3	2	3.06

This is the final data that we have organized after everything. Manually can say that 'agreeableness', 'neuroticism', 'conscientiousness' are the personalities that he should concentrate on for a perfect balance.

Did you answer the questions that you set out to answer as part of your project description and set of experiments?

A. Yes

XIII. References

[1] "five-factor model (FFM)/the OCEAN mode Big 5 Personality test" Bojan Tunguj.

[2] "Myers-Briggs Personality/Top Personality Dataset"Arslan Ali.

[3] K-means clustering - PyTorch API.

[4] "SMOTE for Imbalanced Classification with Python"Dr. Jason Brownlee.

[5] "Shouldering K-Means final model" Timothy Ong.

[6] "Fractal clustering for microarray data analysis" L. Wang. A.Balasubramanian, D. Comaniciu, A. Chakraborty

[7] "Understanding K-means Clustering in Machine Learning"Dr. Michael J. Garbade.